Clase 060 — Model Cards y Responsible ML

Parte: 1 — Machine Learning Clásico · Fuente: Mitchell et al. (2018) Model Cards for Model Reporting + EU AI Act + NIST AI RMF. Duración estimada: 70 min.

Clase 060 — Model Cards y Responsible ML

> *Parte: 1 — Machine Learning Clásico · Fuente: Mitchell et al. (2018) Model Cards for Model Reporting + EU AI Act + NIST AI RMF. Duración estimada: 70 min.*

Objetivo

Aprender a documentar modelos para producción y auditoría: el Model Card (Mitchell et al. 2018, adoptado por Google y luego por la industria) — ficha estandarizada con: propósito, métricas, limitaciones, distribución de datos, riesgos. Conocer el EU AI Act (en vigor 2025-2026), NIST AI RMF, y las plantillas modernas (HuggingFace model cards, Datasheets for Datasets).

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Escribir un Model Card completo con las 9 secciones de Mitchell et al.
- Distinguir un Model Card (sobre el modelo) de un Datasheet (sobre el dataset, Gebru et al. 2018).
- Reportar métricas por subgrupo (no solo global) — clave en fairness.
- Reconocer los 4 tiers de riesgo del EU AI Act (prohibido, alto, transparencia, mínimo).
- Aplicar el NIST AI RMF (Map, Measure, Manage, Govern) en un proyecto real.

Temas

- Secciones de un Model Card: Model Details, Intended Use, Factors, Metrics, Evaluation Data, Training Data, Quant Analyses, Ethical Considerations, Caveats.
- Métricas por subgrupo (sexo, edad, raza, geografía).
- HuggingFace Model Card auto-generation.
- EU AI Act: clasificación de riesgo, obligaciones por tier.
- NIST AI RMF: framework de gestión.
- ISO/IEC 42001 — sistema de gestión de IA.

Definiciones y características

- Model Card: ficha estructurada que documenta un modelo para terceros.
- Datasheet for Datasets: equivalente para datasets — origen, sesgos, demográficos.
- EU AI Act: regulación europea (vigor 2024-2026). Multas hasta 35M€ o 7 % revenue.
- High-risk AI: sistemas en empleo, crédito, educación, justicia, infraestructura. Requieren conformity assessment.
- GPAI (General-Purpose AI): LLMs y similares — obligaciones de transparencia adicionales.
- NIST AI RMF: framework voluntario US, ampliamente adoptado.

Dataset / recursos

- Modelo del proyecto end-to-end (clase 050).
- Plantilla HuggingFace: <https://huggingface.co/docs/hub/model-cards>.

- Librerías: model-card-toolkit (Google).

Ejercicios

1. Model Card básico: para un Random Forest entrenado en California Housing, llenar las 9 secciones. Salvar como MODEL_CARD.md junto al modelo.
2. Subgroup metrics: para un clasificador de credit-g, reportar accuracy y FPR por sex y age_group. Identificar disparidades.
3. Risk classification (EU AI Act): para 5 use cases (recomendación de películas, score crediticio, recurso humano selection, marketing email, detector de spam), clasificar el tier.
4. HuggingFace Card: usar el template de HF; subirla a un repo público si tenés modelo en Hub.
5. NIST RMF: para un proyecto propio, llenar las 4 categorías (Map: contexto, Measure: métricas, Manage: mitigaciones, Govern: ownership).

Homework verificable

Model Card completo para el proyecto end-to-end (clase 050):

1. Las 9 secciones de Mitchell et al. llenadas con datos reales.
2. Métricas por al menos 2 subgrupos demográficos.
3. Sección "Ethical Considerations" con ≥ 3 riesgos identificados y mitigaciones.
4. Sección "Caveats and Recommendations" con limitaciones de validez.

Criterio de aceptación: un revisor externo puede entender propósito, performance, riesgos y cómo usarlo apropiadamente sin acceso al código.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Model Card que solo reporta accuracy globa	Esconde disparidades de subgrupo. Fix: tab
"Intended use" vago ("for predictions")	Inútil. Fix: especificar exactamente qué d
Omitir "Out-of-scope use cases"	No avisás al usuario. Fix: explícito ("NO
Métricas en test, no en producción real	Distribution shift no documentado. Fix: mo
Asumir EU AI Act no aplica	Aplica si el modelo se usa en UE, independ

Preguntas frecuentes

¿Model Card obligatorio?

Por ley: depende de jurisdicción y caso de uso (EU AI Act lo requiere para high-risk). Como buena práctica: siempre.

¿Qué pasa si mi modelo es high-risk EU AI Act?

Obligaciones: conformity assessment, registro en base UE, documentación técnica, supervisión humana, robustez. Multas hasta 7 % revenue.

¿Model Cards en producción industrial?

Sí. Google, Meta, Microsoft, OpenAI, Anthropic — todas publican Model Cards. HuggingFace lo requiere para

modelos en Hub.

¿Datasheet for Datasets equivalente?

Sí, Gebru et al. (2018). Documenta origen, demográficos, sesgos del dataset. HuggingFace tiene Dataset Cards.

¿Y ISO 42001?

Sistema de gestión de IA (publicado dic 2023). Certificación auditable. Análogo a ISO 27001 para seguridad.

Referencias

- Mitchell, M., et al. (2018), Model Cards for Model Reporting, FAT* 2019.
- Gebru, T., et al. (2018), Datasheets for Datasets, CACM 2021.
- EU AI Act texto completo: <https://artificialintelligenceact.eu/>.
- NIST AI Risk Management Framework: <https://www.nist.gov/itl/ai-risk-management-framework>.
- HuggingFace Model Cards: <https://huggingface.co/docs/hub/model-cards>.

Siguiente clase

Clase 061 — CRISP-DM como framework metodológico

Apéndice: notebook (primer bloque)

Implementamos un ModelCard dataclass al estilo Mitchell et al. (2019) y lo rellenamos automáticamente para un LogReg entrenado sobre datos sintéticos con un atributo protegido binario.

```
import json
import numpy as np
import pandas as pd
from dataclasses import dataclass, field, asdict
from typing import Any
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix

rng = np.random.default_rng(42)
np.random.seed(42)
```

Archivos complementarios

- notebook.ipynb